Classification

BA810: Supervised Machine Learning

Nachiketa Sahoo

Recap

- R^2: fraction of variation in Y explained by the regression
- β_k: how much y increases when x_k increases by 1, holding other xs constant
- How reliable is the estimate?
  - Confidence interval around $\hat{\beta}_k$
  - p-value of $\hat{\beta}_k$: probability of obtaining this value by chance even if true β_k was 0
- Overfitting (low training error, high test error) and underfitting (high train and test error)
- The bias-variance tradeoff: predict at a point refitting to different training sets
  - Bias: Average prediction is off target
  - Variance: Predictions varying a lot
- Bias-variance tradeoff in practice
  - Complex models have more variance, but less bias (more likely to overfit)
  - Pick complexity to minimize validation data error

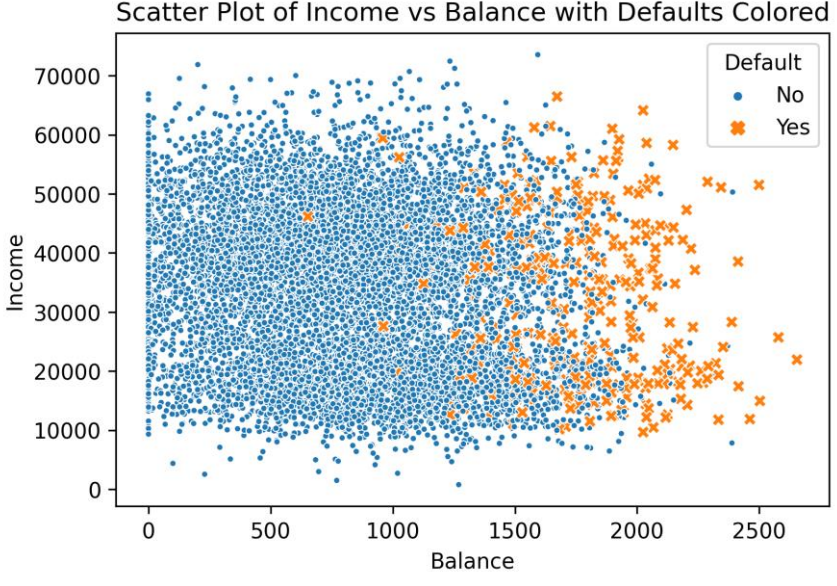
Classification

- Qualitative variables take values in an unordered set C:

$$eye\ color \in \{brown, blue, green\}$$
$$email \in \{spam, good\}$$

- Given a feature vector X and a qualitative response Y taking values in the set C, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e., $C(X) \in C$.
  - Often, we are interested in estimating the probabilities that X belongs to each category in C
  - Matters if predicting spam with 90% probability of 51%
  - Thus, a continuous score is typically estimated

Example: Credit Card Default



Can we use Linear Regression?

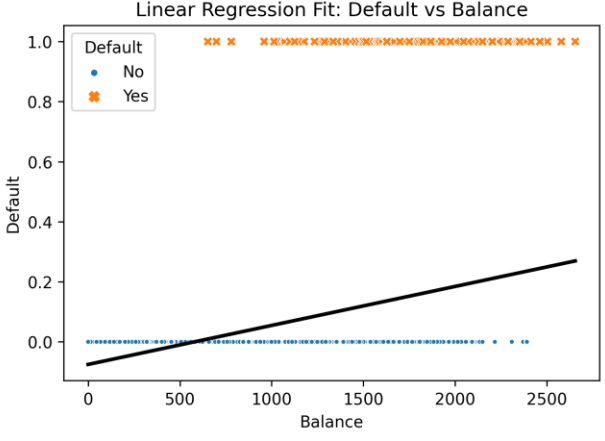
- Suppose for the Default classification task that we code

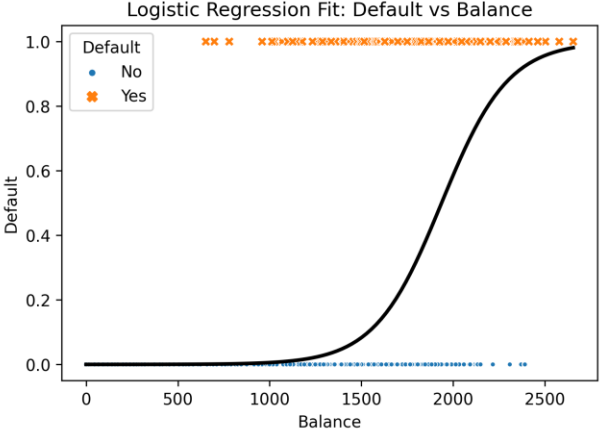
$$Y = \begin{cases} 0 \text{ if No} \\ 1 \text{ if Yes} \end{cases}$$

and take them to mean probability of default.

- Can we simply regress of Y on X and classify as Yes if $\hat{Y} > 0.5$?

Linear vs Logistic Regression





Logistic Regression

Let $p(X) = \Pr(Y = 1|X)$. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}, \text{ where } e \approx 2.7 \text{ [Euler's number.]}$$

βs are chosen to maximize the probability of observed classes in the training data.

$p(X)$, per this formula, must be between 0 and 1.

Rearranging gives:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

log odds ratio or logit transformation of $p(X)$ is a linear function of features/columns.

Linear versus Logistic Regression

Interpreting the coefficients

- Linear regression: $y = \beta_0 + \beta_1 x$
  - When x increases by 1, y increases by β_1

- Logistic regression: $\log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right) = \beta_0 + \beta_1 x$, or $\frac{p(y=1|x)}{1-p(y=1|x)} = e^{\beta_0+\beta_1 x}$
  - Increase x by 1 → RHS: $e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x} \times e^{\beta_1}$
  - For small β_1s: $e^{\beta_1} \approx 1 + \beta_1$
  - when x increases by 1, the odds ratio of positive class ($y = 1$) increases by β_1 fraction

Making Predictions

Let's say the estimates are

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Estimated probability of default at balance = $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006: \text{No}$$

With balance = $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586: \text{Yes}$$

Linear Regression vs. Logistic Regression continued

Target has >2 values

A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

- Can we use a linear regression here?

- **Multiclass Logistic Regression** (also known as: **multinomial regression**) or other classifiers are more appropriate.

$$P(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}{\sum_l^K e^{\beta_{0l}+\beta_{1l}X_1+\cdots+\beta_{pl}X_p}}$$

A linear function for *each* class.

Bayes Theorem : A Way to Invert Conditional Probabilities

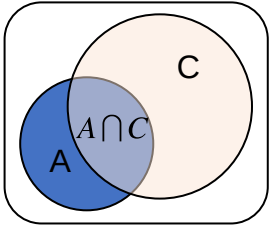
- Let A and C be two binary random variables
  - Each observation is a joint observation of two
- Conditional Probability

$$P(C|A) = \frac{P(A,C)}{P(A)} \text{ and } P(A|C) = \frac{P(A,C)}{P(C)}$$

- Bayes theorem

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

Use easier to collect conditional probability
to estimate the hard to collect one



Example of Bayes Theorem

- Consider the following scenario
  - A doctor knows that 50% of the patients with meningitis have stiff neck
  - Probability of any patient having meningitis is 1/50,000
  - Probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

A simple Bayesian classifier

More Complex Classifiers using Bayes Rule

- Let Y be the target class taking K values and X be the *set* of predictors, then

$$P(Y = k|X = x) = \frac{\Pr(X = x|Y = k)\Pr(Y = k)}{\Pr(X = x)}$$

- $\Pr(Y = k)$ is the marginal or prior probability of class k (π_k)
- $\Pr(X = x|Y = k)$ is the distribution of the attribute in class
  - Both can be calculated from the training data
- $\Pr(X = x)$ can be calculated as $\sum_{l=1}^{K} \Pr(X = x|Y = l)\Pr(Y = l)$.
  - Though often unnecessary for predicting class.
- Naïve Bayes classifier
  - Components of X are conditionally independent of each other given (within) a class
    - Strong assumption, works OK for prediction, though probabilities are not accurate ("well calibrated")
  - Some components can be categorical and others numeric (with different distributions)

K-nearest Neighbor (k-NN)

- To predict the class label of each item in test set
  - find the k most similar items in training set
  - use the label that is most common among them
- How do you select k?
  - Try different k values and see what leads to lowest error in validation data
  - Set aside a portion as test data
  - Using the remaining training data use cross validation to choose
- Illustration

- Any challenges with this approach?

Evaluating a Classifier

Metrics for Performance Evaluation

- We focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- **Confusion Matrix** (cells contain counts of records with certain predicted and actual class labels)

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

The correctness
Of the prediction

In two class setting:

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

The prediction

Metrics for Performance Evaluation…

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Most widely-used metric:

$$Accuracy = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

- When you have more than two classes
  - sum of diagonal numbers / sum of all numbers

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class **No** examples = 9990
  - Number of Class **Yes** examples = 10
- If model predicts everything to be class **No**, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class **Yes** example
- Many real-world datasets have skewed class distribution
  - E.g., fraudulent credit card transactions aren't as common as good transactions
- Better alternative: Balanced Accuracy
  - Average of fractions of true Yes's predicted to be Yes and true No's predicted to be No

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$$

  - Or the average of the positive and negative recall (defined next)

Evaluation from Retrieval Perspective

How good is the classifier in getting/detecting the positive records?

Can it do so without making many mistakes?

- $Precision\ (p) = \frac{a}{a+c}$

How many of the positive records can it get?

- $Recall(r) = \frac{a}{a+b}$
  - Possible to rig these scores. How?
- $F-measure(F) = \frac{2rp}{r+p}$

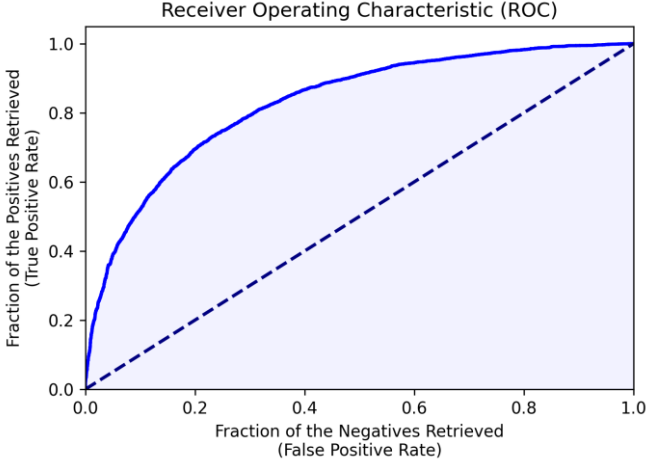
F is the Harmonic Mean of Precision and Recall

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{p} + \frac{1}{r}\right)$$

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

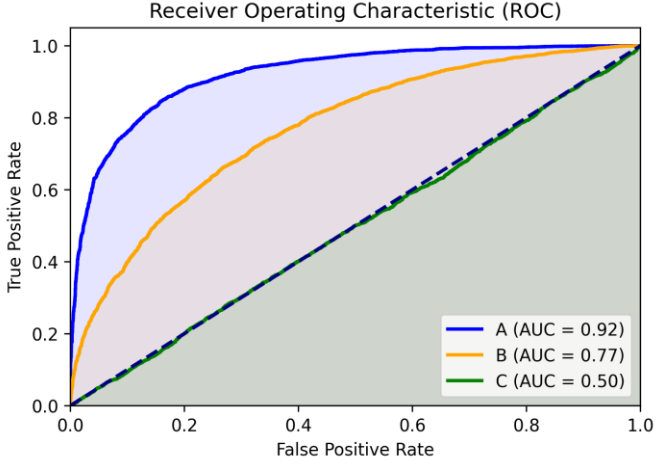
Receiver Operating Characteristic (ROC) Curve

- Most classifiers produce a continuous score of a record being positive
  - The larger this score, the record is more likely to be positive
- Asked to find one record that is positive
  - which record would you present?
- A test dataset sorted in the decreasing order of probability of records being positive (as estimated by the classifier)
  - A classifier asked to find more and more positive records from the test set
  - How would the performance evolve?



ROC Curve

Which is the best performing model in the following three graphed?



Summary

Classifiers

1. Logistic regression: fits a linear function to odds ratio
2. Bayes classifiers: use Bayes rule to invert the direction of conditional probability
3. K-nearest-neighbors: predict the label frequent nearby

1&2: model based (learn and use compact statistics from training data)

3: instance based (remember the whole training data and use to predict)

Evaluation

- Confusion matrix,
  - True Positive, True Negative, False Positive, False Negative
- (Balanced) Accuracy, Precision, Recall, F-measure
- ROC curve, Area Under the Curve (AUC)

Next Class

- Cost based evaluation, Cross validation, Regularization